# Optimization of Cross-Lingual Voice Conversion With Linguistics Losses to Reduce Foreign Accents

Yi Zhou, *Member, IEEE*, Zhizheng Wu, *Senior Member, IEEE*, Xiaohai Tian, *Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Cross-lingual voice conversion (XVC) transforms the speaker identity of a source speaker to that of a target speaker who speaks a different language. Due to the intrinsic differences between languages, the converted speech may carry an unwanted foreign accent. In this paper, we first investigate the intelligibility of the converted speech and confirm the performance degradation caused by the accent/intelligibility issue. With the goal of generating native-sounding speech, this paper further proposes a novel training scheme with two additional linguistic losses for speech waveform generation: 1) a frame-wise phonetic content loss derived from bottleneck features, and 2) an automatic speech recognition loss on characters. Experiments were conducted between English and Mandarin Chinese conversions. The experimental results confirmed that the generated speech sounds more natural with the proposed linguistic losses and the proposed solution significantly improves speech intelligibility.**

***Index Terms*—Cross-lingual voice conversion (XVC), speech intelligibility, linguistic loss.**

## I. Introduction

CROSS-LINGUAL Voice Conversion (XVC) seeks to change the speaker identity of a speech sample from one (source speaker) to another (target speaker), under the constraint that the target is a foreign speaker [1]. For example, we would like to make Albert Einstein to speak perfect Chinese. XVC techniques can be used for a number of applications such as foreign language education [2], speech translation [3], and foreign movie dubbing [4], [5]. The XVC studies have shown that we are able to preserve very well the speech content of the source, however, we inevitably introduce a foreign accent to the converted speech in the process due to the intrinsic differences between the languages [6], [7], [8], [9].

Over the past decades, a number of XVC techniques have been investigated. In general, they mainly focus on three different aspects: 1) improving the source-target speech alignment for parallel training, 2) employing advanced models for higher audible quality, and 3) advancing XVC performance with limited training data. In the early studies, voice conversion depends on parallel data for model training [1], [2]. With the development of more advanced acoustic model from speech recognition [10], [11], [12], [13] and the advent of deep learning, we are now able to train a conversion function with non-parallel data [14], [15], [16], [17]. Most of the studies focus on a better conversion function [17], [18], [19], [20], [21], [22], [23] and take care of different aspects of speech, such as accent [24], prosody [25] and emotion [26], [27].

Despite much progress, speech intelligibility in XVC remains a challenge. There are but a few studies that focus on speech intelligibility [28], [29]. Each spoken language has its unique phonetic and prosodic characteristics. The phonetic inventory, phonotactic rules, stress patterns, and rhythm are considerably different between languages [30]. The differences between the source and target languages in XVC may result in a non-native accent in the converted speech, which adversely affects the intelligibility of the converted speech [31].

In this work, we first investigate the speech intelligibility problem (i.e., foreign accent) in the converted speech from the state-of-the-art XVC techniques. We then propose a novel technique to alleviate the foreign accents. This is achieved by optimizing the XVC conversion model with two additional linguistic losses, namely bottleneck feature (BNF) loss and automatic speech recognition (ASR) loss. The proposed XVC framework takes the BNFs as input and generates speech waveform as the output.

Specifically, a BNF is a time-versus-class matrix representing the phonetic content for each specific time frame [32], which can be obtained from a pre-trained speech recognition neural network. Therefore, the BNF loss is a frame-wise loss computed between the input BNFs and the predicted BNFs extracted from the generated waveform. On the other hand, we define the ASR

Manuscript received 17 April 2022; revised 8 November 2022; accepted 11 April 2023. Date of publication 27 April 2023; date of current version 19 May 2023. This work was supported by in part by the National Natural Science Foundation of China under Grant 62271432, and in part by the Human-Robot Collaborative AI for Advanced Manufacturing and Engineering under Grant A18A2b0046, Agency for Science, Technology and Research, Singapore. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sakriani Sakti. *(Corresponding author: Zhizheng Wu.)*

Yi Zhou is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: e0154158@u.nus.edu).

Zhizheng Wu is with the Shenzhen Research Institute of Big Data, School of Data Science, Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: wuzhizheng@cuhk.edu.cn).

Xiaohai Tian is with the Speech and Audio Department, Bytedance AI lab, Singapore 569933 (e-mail: xiaohai171@gmail.com).

Haizhou Li is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077, and also with the Shenzhen Research Institute of Big Data, School of Data Science, Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: haizhou.li@u.nus.edu).

Digital Object Identifier 10.1109/TASLP.2023.3271107

loss as the error between the ground-truth character sequence and the ASR transcript derived from the generated waveform. The linguistic losses are expected to preserve the linguistic content therefore maintain the speech intelligibility during the voice conversion.

The rest of this paper is organized as follows. In Section II, we present a study on speech intelligibility to motivate this work. In Section III, we propose the linguistic losses for XVC model training and present the overall framework. In Section IV, we perform XVC experiments between English and Mandarin Chinese to validate the proposed idea. Finally, Section V concludes this paper.

## II. Motivation

Speech intelligibility measures how well the linguistic content is delivered via the speech signal, which is a key attribute of speech quality. In a XVC task, speech intelligibility can be adversely affected by unwanted accent. According to the prior work [6], [7], [8], [9] and the large-scale shared tasks as part of the Voice Conversion Challenge 2020 (VCC 2020) [33], it is apparent that there is a quality gap between intra-lingual and cross-lingual voice conversion [33], [34], [35], [36], [37], [38]. However, to the best of our knowledge, there is no reported inquiry into the issue yet as to how to quantify such a quality gap. It is hypothesized that the foreign accent arising from XVC leads to a speech intelligibility drop [6], [7], [8], [9].

We start by evaluating the submissions to VCC 2020 to validate the hypothesis. Then, we evaluate the converted samples from our in-house XVC system, which was built from bilingual data. In addition, we report the speech recognition performance in terms of word error rate (WER) or character error rate (CER) as a proxy of speech intelligibility. We also report the XAB preference test in a subjective evaluation.[1]

### A. Speech Intelligibility: VCC 2020 System Submissions

VCC 2020 is the most recent large-scale international benchmarking, that includes both intra-lingual and cross-lingual voice conversion tasks. In total, 33 systems with different configurations were submitted to VCC 2022. Based on the evaluation results in terms of speech naturalness and speaker similarity, the top four systems, e.g., T10 [39], T13 [40], T25 [41], and T29 [42], were selected to assess speech intelligibility of both intra-lingual and cross-lingual conversion. English utterances from four English source speakers (SEF1, SEF2, SEM1, SEM2) were chosen for evaluation. For the intra-lingual voice conversion task, English speakers TEF1 and TEM1 were selected as the target speakers, while two Mandarin target speakers (TMF1, TMM1) were selected for the cross-lingual task. Each source speaker has 25 utterances, which means that each selected system contributes $(4 \times 25 \times 2) + (4 \times 25 \times 2) = 400$ converted speech samples.

- *WER:* In VCC 2020, English is the source speech for both the intra-lingual and cross-lingual conversion tasks. We used the Google Cloud Speech-to-Text[2] service to transcribe speech samples. The English (United States) language was selected from the Google API, which was trained with a huge amount of English data of native speakers. The WERs were reported on the transcription. As shown in Table I, the speech samples from cross-lingual conversion show a higher WER than those from intra-lingual conversion across all the four systems. In addition, we note that the average WERs of the intra-lingual and cross-lingual voice conversions are 18.37 % and 23.57 %, respectively. This suggests that the converted samples in the cross-lingual task are less intelligible than the intra-lingual ones.
- *XAB Preference Test:* In the XAB test, we invited listeners to judge whether the converted speech sounded as native as the source in terms of pronunciation. X was the natural source speech that served as the reference, while A and B were the converted speech samples from two different systems. X, A, and B were of the same content. The participants listened to the reference first, then to samples A and B in random order. They were instructed to compare the intonation, articulation, and pronunciation at both the word and sentence levels with the reference. 20 listeners with professional English proficiency participated in this listening test, and each listener evaluated 20 utterance pairs. As shown in Fig. 1, the listeners consistently prefer intra-lingual outputs than cross-lingual ones over intelligibility across all systems.

TABLE I
Average WER (%) of the Converted Speech of Four Top Ranking Systems From VCC 2020 and the Source Natural Speech. Each System Produces Both Intra-Lingual and Cross-Lingual Conversion Samples

| | Source | T10 | T13 | T25 | T29 | Average |
|---|---|---|---|---|---|---|
| Intra-Lingual | 13.79 | 11.26 | 18.69 | 20.19 | 23.33 | 18.37 |
| Cross-Lingual | | 15.11 | 22.99 | 24.68 | 31.48 | 23.57 |

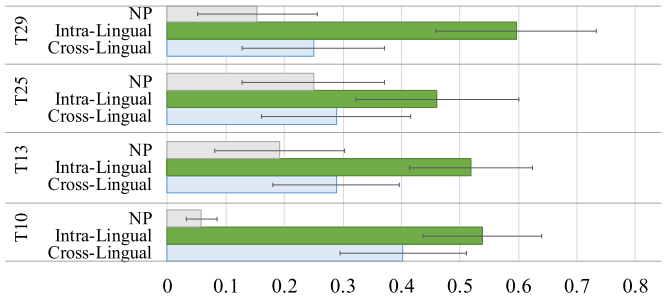


Fig. 1. Speech intelligibility listening test results of the converted speech with 95% confidence intervals in intra-lingual and cross-lingual voice conversion tasks of the four top ranking systems from VCC 2020.

### B. Speech Intelligibility: In-House XVC System

We further studied the speech intelligibility with our in-house XVC system [37]. We selected 2 source speakers (MF6, MM6) and 2 target speakers (MF7, MM7) from the EMIME database [43] for both intra-lingual and cross-lingual conversions. All speakers are English-Mandarin bilingual. Our in-house XVC system consists of a feature conversion model and

[1] Samples could be found at https://vcsamples.github.io/xvc_study
[2] https://cloud.google.com/speech-to-text, accessed on 08/09/2021.

TABLE II
THE AVERAGE WER (%) AND CER (%) OF THE CONVERTED ENGLISH AND MANDARIN SPEECH IN INTRA-LINGUAL VOICE CONVERSION AND CROSS-LINGUAL VOICE CONVERSION TASKS FROM OUR OWN XVC SYSTEM. SOURCE IS THE NATURAL SOURCE SPEECH

| Voice Conversion | ENG WER (%) | MAN CER (%) | Average |
|---|---|---|---|
| Source | 14.61 | 12.11 | 13.36 |
| Intra-Lingual | 24.01 | 21.68 | 22.85 |
| Cross-Lingual | 35.66 | 29.87 | 32.77 |

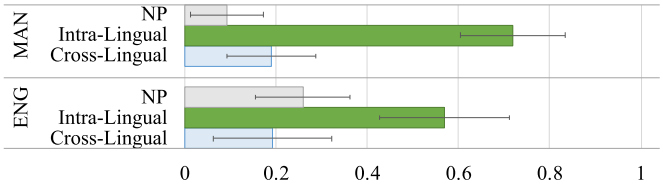


Fig. 2. Speech intelligibility listening test results of the converted speech with 95% confidence intervals in intra-lingual voice conversion and cross-lingual voice conversion tasks from our own XVC system.

a WaveRNN [44] vocoder. The conversion model maps input BNF to spectral features using two 256-dimensional BLSTM layers [37].

For each target speaker, we trained two conversion models. One English model was trained with 145 English utterances and the other Mandarin model with 145 Mandarin utterances. In total, there are 4 conversion models with the same network configurations. To perform intra-lingual voice conversion, the source utterances and the conversion models are in the same language. For example, English utterances from the source speakers are converted by the English conversion models. For cross-lingual voice conversion, the languages of the source speech samples and the conversion models are different. For each source speaker, 25 English utterances and 25 Mandarin utterances were converted into both target speakers. In total, there are $2 \times 2 \times 25 \times 2 = 200$ English and $2 \times 2 \times 25 \times 2 = 200$ Mandarin converted utterances.

- *WER/CER:* English and Mandarin speech were transcribed by Google API and iFLYTEK Open Platform[3], respectively. The WER values were reported for English speech while the CER scores were calculated for Mandarin speech. The results are summarized in Table II. For English, intra-lingual achieves a WER of 24.01%, while the WER of cross-lingual is only 35.66%. Similarly, the Mandarin CER value of intra-lingual and cross-lingual are 21.68% and 29.87%, respectively. The average WER value for intra-lingual is 22.85%, while it increases to 32.77% for cross-lingual. The error rate increment in cross-lingual task indicates the problem of speech intelligibility degradation in XVC.
- *Speech Intelligibility XAB Test:* The same XAB test was also conducted for speech intelligibility. Fig. 2 presents the test results according to the generated speech language. It can be seen that the intra-lingual task significantly outperforms the cross-lingual counterpart in both scenarios of target language with English and Mandarin. It demonstrates the performance gap between XVC and intra-lingual voice conversion, which also highlights the speech intelligibility problem caused by the different languages in XVC. Moreover, the degradation of intelligibility is more pronounced in Mandarin.

[3]https://www.iflyrec.com/, accessed on 11/10/2021

### C. Observations

In the speech intelligibility experiments with the VCC 2020 submissions and our in-house XVC systems, we have the following findings,

- There is a significant intelligibility gap between intra-lingual and cross-lingual conversions, even when using the same bilingual speaker's data as the target (as seen in our in-house system).
- The intelligibility of Mandarin Chinese, a tonal language, is degraded more than that of English in the cross-lingual conversion task.
- The intelligibility gap is independent of system implementation. We have same observations across the board.

From these findings, we argue that the intelligibility gap is not due to system implementation or speaker variation, but rather the results of phonetic and prosodic differences between languages. During the training of XVC systems, the mean squared error (MSE) is usually applied to minimize the error between the predicted and the ground-truth Mel-spectrogram at the frame level. However, the phonetic and prosodic differences are manifested at a segmental or utterance level. To maintain the intelligibility, we introduce two linguistic losses, which seek to preserve the content, and the accent of the source speech at the same time. This will be the main focus of this paper.

## III. LINGUISTIC LOSSES

The XVC network is expected to preserve the input phonetic content, at the same time, produce highly intelligible speech. We propose to incorporate two additional linguistic losses as part of the training objective, i.e. a BNF loss and an ASR loss. The BNF loss is introduced to maintain the phonetic content at the frame level, while the ASR loss serves as a perceptual loss [45] to ensure intelligibility at the word level. Let us start by describing the BNF-to-Waveform XVC framework. Then, we introduce the proposed BNF loss and ASR loss for foreign accent reduction. Finally, we describe the XVC framework with two proposed linguistic losses on the speech waveform.

### A. BNF-to-Waveform XVC Framework

BNF is the bottleneck feature that is believed to represent the linguistic information. We use the feature frames before the last softmax output layer of an ASR model as the BNF features. The ASR model is pretrained on a separate speech recognition dataset. The BNF-based method is widely used in recent voice conversion studies [33]. By adopting BNF as inputs, the conversion model learns a feature mapping between linguistic features

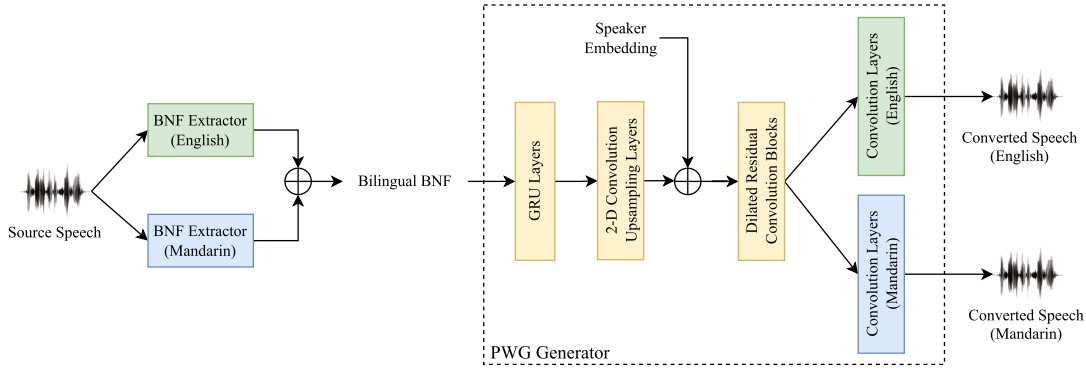


Fig. 3. The neural architecture of the BNF-to-Waveform cross-lingual voice conversion model at run-time. We employ two language-specific output layers for two output languages. The $\bigoplus$ symbol denotes a concatenation operation.

and the corresponding acoustic features or speech waveform. We have studied a BNF-to-Waveform end-to-end architecture for XVC, that directly predicts waveform from BNF [46]. The BNF-to-Waveform framework achieves high performance for both speech quality and speaker similarity. Therefore, it is taken as the XVC network in this study.

The BNF-to-Waveform framework is a modification of the Parallel WaveGAN (PWG) network [47] consisting of a generator, a discriminator and multi-resolution blocks. In particular, its generator is modified to take bilingual BNFs as input and generate English and Mandarin speech waveform separately via multi-task learning. Bilingual BNF is previously proposed for XVC [37], which is formed by stacking two monolingual BNF vectors extracted from two automatic speech recognition (ASR) systems trained in two languages, respectively. It provides linguistic representation to characterize two phonetic language systems in XVC tasks [37]. Moreover, the model training with language-specific layers is another of our previous studies on XVC, where the conversion model has a hidden section and a output section [46], [48]. The hidden sections is shared between two languages, while the output section has two language-specific heads, each for one language [48]. The shared layers serve as the bridge between the two languages, while the language-specific layers take care of the acoustic renderings individually. In this way, the conversion model has gained the capability to generate high-quality speech of specific languages [48].

The rest components are kept the same as the original PWG network [47]. Fig. 3 presents the run-time BNF-to-Waveform framework for XVC. The bilingual BNF is implemented by concatenating BNFs from English and Mandarin BNF extractors. It is then passed through GRU layers and 2-D convolutional upsampling layers. The upsampled features are concatenated with speaker embeddings and then pass through the dilated residual convolution blocks in the corresponding language. There are two language-specific convolution layers, one for English and the other for Mandarin. The overall loss function in the PWG network $\mathcal{L}_{PWG}$ contains three different losses: the multi-resolution STFT loss, the adversarial loss, and the discriminator loss [47]. Speaker embedding is extracted from a pre-trained speaker embedding extractor, which has been omitted from Fig. 3.

During training, we update the the language-specific convolution layers only with the data from the corresponding language. The English and Mandarin BNF extractors are pre-trained models whose weights are frozen during training. The remaining network blocks, including the GRU layers, upsampling layers, and dilated residual convolution blocks, are updated in a language-independent fashion.

At run-time, the source speech is passed to the English and Mandarin BNF extractors to obtain bilingual BNF, while the target speech is used for speaker embedding extraction. The bilingual BNF is the input to the trained PWG generator, and the waveform output from the desired convolution layers in the same language as the source speech is the converted speech.

### B. Proposed Linguistic Losses

We propose two additional losses for intelligibility, namely BNF loss and ASR loss. Since both BNF loss and ASR loss evaluate the errors associated with the linguistic information, the two losses are imposed on the network to ensure the proper delivery of the linguistic content.

*1) Frame-Wise BNF Loss:* The BNF loss is implemented to preserve the phonetic content of the input during model training, that is calculated at the frame level between the BNFs extracted from the input speech and the corresponding reconstructed speech following the equation:

$$\mathcal{L}_{BNF} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{b}_i - \hat{\mathbf{b}}_i)^2, \tag{1}$$

where $N$ is the total number of frames. $\mathbf{b}_i$ and $\hat{\mathbf{b}}_i$ denote the BNFs extracted from the input speech and the reconstructed speech at the $i$-th frame, respectively.

Since BNF is assumed to be a speaker-independent linguistic representation [49], the BNF loss is expected to encourage the conversion model to generate speech waveform of the same phonetic content as the natural input speech, and discourage phonetic variation.

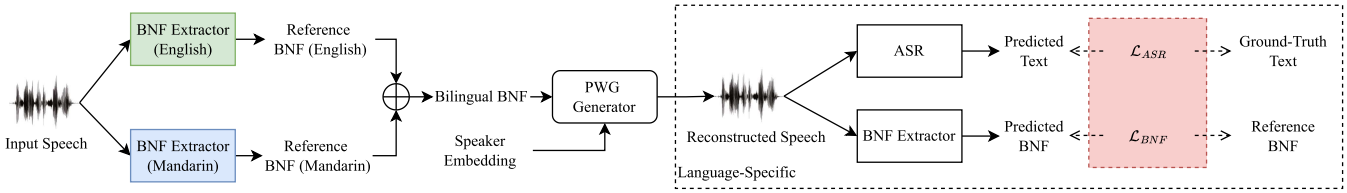


Fig. 4. The training diagram of the XVC framework with the proposed linguistic losses. $\mathcal{L}_{ASR}$ and $\mathcal{L}_{BNF}$ denote the segmental ASR loss and the frame-wise BNF loss respectively. The reference BNF is extracted from the input speech, while the ground-truth text is the transcript of the input speech. Linguistic losses are calculated on a language-specific basis.

*2) Segmental ASR Loss:* The ASR loss reflects the human perception of speech intelligibility. We adopt the speech recognition error derived from the converted speech as the proxy of speech intelligibility. In this study, a pre-trained end-to-end ASR model [50] is used. The ASR model has a unified two-pass joint connectionist temporal classification (CTC) decoder and attention-based encoder-decoder (AED) decoder architecture. The total ASR loss is obtained by combining the CTC loss and AED loss:

$$\mathcal{L}_{ASR} = \lambda \mathcal{L}_{CTC}(\mathbf{x}, \mathbf{y}) + (1 - \lambda)\mathcal{L}_{AED}(\mathbf{x}, \mathbf{y}), \qquad (2)$$

where $\mathbf{x}$ is the spectral feature, $\mathbf{y}$ is the corresponding character label, $\mathcal{L}_{CTC}(\mathbf{x}, \mathbf{y})$ and $\mathcal{L}_{AED}(\mathbf{x}, \mathbf{y})$ are the CTC and AED loss respectively while $\lambda$ is the coefficient to balance the CTC and AED losses [50]. For English, the character labels consist of the 26 alphabets and 4 symbols, while for Mandarin Chinese, 4,230 Chinese characters and 3 symbols are used as the character labels.

Since the English and Mandarin Chinese ASR models are trained independently, the resulting ASR models are supposed to model the respective phonetic systems. Therefore, the ASR loss from such pre-trained models serves as the perceptual loss [45] to ensure the native accent expected out of the converted speech. It is worth noting that the ASR loss is introduced at the character level, that is complementary to the BNF loss at the frame level.

### C. XVC With Linguistic Losses on Speech Waveform

We integrate the two proposed linguistic losses into the BNF-to-Waveform framework to optimize the conversion model jointly with other losses. Fig. 4 presents the overall architecture of the XVC system. Note that the BNF extractor and ASR networks are pre-trained neural networks that are differentiable during back-propagation.

During training, the BNF extractors first extract the reference BNFs from the input speech and concatenates them as bilingual BNF, which is utilized as input to the BNF generator. The BNF generator reconstructs the speech waveform directly from the bilingual BNF conditioned on speaker embeddings. Subsequently, the ASR system and the BNF extractor obtain the predicted text and BNF from the reconstructed speech waveform in the same language. The ASR loss is the error between the predicted and the ground-truth character sequences, while the BNF loss is the difference between the predicted BNFs and the reference BNFs. In this way, the XVC network is optimized by the combined PWG network loss $\mathcal{L}_{PWG}$, BNF loss $\mathcal{L}_{BNF}$, and ASR loss $\mathcal{L}_{ASR}$. The overall loss follows the equation below:

$$\mathcal{L}_{Total} = \mathcal{L}_{PWG} + \lambda_{BNF}\mathcal{L}_{BNF} + \lambda_{ASR}\mathcal{L}_{ASR}, \qquad (3)$$

where $\lambda_{BNF}$ and $\lambda_{ASR}$ $(0 < \lambda_{BNF} < 1,\ 0 < \lambda_{ASR} < 1)$ are the coefficients for BNF loss and ASR loss, respectively. During training, we only update the parameters in the PWG generator. The BNF extractor and the ASR system are not involved in the training. The conversion process at run-time is exactly the same as described in Section III-A.

## IV. Experiments

We validated the proposed XVC framework with linguistic losses by converting between English and Mandarin Chinese. This section presents the experimental details.

### A. Database and Data Processing

Multiple corpora are used to pre-train the BNF extractors, speech recognition models, and XVC models. All speech signals are re-sampled to 16 kHz. The 80-dimensional Mel-spectrogram is extracted with a window size of 50 ms and a hop size of 12.5 ms using Torchaudio. It is used as input features for the BNF extractors, the ASR systems, and the speaker embedding network.

*1) BNF Extractors:* Two BNF extractors are pre-trained for English and Mandarin, respectively. The English BNF extractor is trained using the 460-hour Librispeech database [51]. The BNF extractor for Mandarin is trained with multiple Mandarin corpora containing 1,238 hours of speech. The corpora included AIDataTang [52], AISHELL-1 [53], MagicData [54], PrimeWords [55], ST-CMDS [56], and THCHS-30 [57].

Both the English and Mandarin models have the same network architecture, which consists of three $5 \times 1$ convolutional layers. Each 512-channel convolutional layer is followed by batch normalization and ReLU activation. The output is then passed through three 512-unit BLSTM layers, followed by a dense 256-unit bottleneck layer that output BNF features at run-time. The final softmax layer consists of 5,808 and 9,864 units for English and Mandarin, respectively. The output is the senone class. The frame accuracy of the English senone classification is 73.84%, and is 76.37% for Mandarin. The BNFs for English and Mandarin are both 256-dimensional.

*2) ASR Systems:* The English and Mandarin ASR systems are pre-trained with the same network architecture: a conformer-based [58] encoder and an attention-based decoder. The English ASR is trained using the Librispeech database (460 hours) [51],

TABLE III
THE SOURCE-TARGET SPEAKER PAIRS IN THE XVC EXPERIMENTS. ALL SPEAKERS ARE ENGLISH AND MANDARIN BILINGUAL. → INDICATES THE XVC CONVERSION DIRECTION

| Source-Target Speaker Pairing |
|---|
| MF2 → MM2 (Female → Male) |
| MF4 → MF2 (Female → Female) |
| MM1 → MF4 (Male → Female) |
| MM2 → MM1 (Male → Male) |

while the Mandarin ASR is trained with AISHELL-1 (151 hours) [53]. The WeNet toolkit [50] is utilized for both English and Mandarin ASR training. The output is the text transcription (i.e., word sequence). The model configuration follows that in [50].

The English ASR achieves WER results of 5.31% and 10.12% on LibriSpeech clean and other test datasets [51]. The Mandarin ASR test CER on the AISHELL-1 [53] test set is 5.72%. The WER/CER is comparable to the published WeNet baseline system [50].

*3) Speaker Embedding:* The speaker embedding extractor is a language-agnostic network [59]. It is pre-trained on the AISHELL-2 database [60] and then fine-tuned with English and Mandarin speech data from 100 speakers. The network configuration and data description can be found in [59].The extracted speaker embedding dimension is 256.

*4) XVC Database:* The conversion models in the XVC experiments are trained with both English and Mandarin monolingual speech data from 100 speakers (50 English speakers and 50 Mandarin speakers), with each speaker contributing 150 utterances. 50 English speakers are randomly selected from the VCTK database [61], while 50 Mandarin speakers are in the Data-Baker Mandarin Library.[4] During testing, 4 bilingual speakers (MF2, MF4, MM1, MM2) are selected from the EMIME database [43] for conversion, with each speaker contributing 20 English utterances and 20 Mandarin utterances. We convert between the 4 bilingual speakers, resulting in 80 ($20 \times 4 = 80$) converted English utterances and 80 ($20 \times 4 = 80$) converted Mandarin utterances. The details of the speaker pairs can be found in Table III.

### B. XVC Systems

We implemented 7 different XVC systems for comparison. They are variants of two model architectures: BNF2Mel and BNF2Wav architectures. The BNF2Mel architecture converts the BNF features of the source to the Mel-spectrogram of the target and uses a neural vocoder to reconstruct the speech waveform, whereas the BNF2Wav conversion model converts the BNF features of the source directly into the target waveform. The detailed experimental system configurations are summarized in Table IV.

*1) BNF2Wav:* This serves as the baseline system that converts input BNFs directly into a speech waveform. The input features are 512-dimensional bilingual BNFs, i.e. a combination of English and Mandarin BNFs. The output features are the 16-bit speech signals. As shown in Fig. 3, two GRU layers have 512 hidden units with an upsampling factor of [2, 2, 5, 10]. Both the English and Mandarin convolutional layers have the same model hyper-parameters, which consist of two $1 \times 1$ convolutional layers. The batch size is 8. All the other settings are the same as in the original PWG work [47].

*2) BNF2Wav($F0$):* This is an extension to BNF2Wav by taking the fundamental frequency $F0$ as an additional input. During training, the WORLD vocoder is used to obtain $F0$ from the input speech, which is then concatenated to the bilingual BNFs. At run-time, $F0$ is first extracted from the source speech by the WORLD vocoder, and then linearly transformed to augment the BNFs. As in many intra-lingual and cross-lingual implementations of voice conversion, the fundamental frequency is often used to control prosody [33]. With this model, we would like to know the effect of the an explicit $F0$ input on speech intelligibility.

*3) BNF2Mel-BNF:* We would like to explore the effect of an additional frame-wise BNF loss [41], [62] on a standard BNF2Mel network architecture, which converts BNFs into Mel-spectrogram, and reconstructs the speech waveform using the PWG vocoder. This model is adapted from the T25 system [41] in VCC 2020 [33]. It takes 512-dimensional bilingual BNFs as the input, and generates 80-dimensional Mel-spectrogram as the output. To apply the BNF loss, we obtain the output BNFs from the Mel-spectrogram generated by the language-specific output layers [62]. The hyper-parameters follow those in [62]. The architecture of the spectral conversion network is summarized in Table V. The PWG neural vocoder [47] is pre-trained with the VCTK corpus [61]. It takes the Mel-spectrogram as input to predict 16-bit speech signals at 16 kHz. As only $\lambda_{BNF}$ in (3) is applied, we empirically set $\lambda_{BNF} = 0.0015$ and $\lambda_{ASR} = 0.0$.

*4) BNF2Wav-BNF:* This system is similar to the BNF2Mel-BNF system except that we generate the speech waveform directly using the PWG generator shown in Fig. 3, without involving Mel-spectrogram.

*5) BNF2Wav-ASR:* This system is similar to the BNF2Wav-BNF system except that the proposed ASR loss is used instead of BNF loss. In other words, we empirically set $\lambda_{ASR} = 0.006$ and $\lambda_{BNF} = 0.0$ in (3).

*6) BNF2Wav-BNF-ASR:* This is the proposed XVC system with two linguistic losses. As shown in Fig. 4, both pre-trained BNF extractors and ASR systems are used to extract predicted BNFs and text, respectively, from the reconstructed speech waveform. The coefficients $\lambda_{BNF}$ and $\lambda_{ASR}$ for the BNF loss and ASR loss in (3) are 0.008 and 0.002, respectively.

*7) BNF2Wav-BNF-ASR($F0$):* This system is similar to the BNF2Wav-BNF-ASR system, except that it adds the additional feature $F0$ to the bilingual BNFs as input features for the conversion model. All settings are kept the same as the BNF2Wav-BNF-ASR system.

To find the optimal combination of the coefficients $\lambda_{BNF}$ and $\lambda_{ASR}$ for the BNF loss and ASR loss. We found that the

[4]http://www.data-baker.com/hc_pm_en.html

TABLE IV
SUMMARY OF THE EXPERIMENTAL SYSTEMS, THEIR CONFIGURATIONS, AND THE RESULTS OF THE OBJECTIVE MCD (DB), RMSE (DB) AND WER/CER (%) VALUES. THE $\bigoplus$ SYMBOL DENOTES THE CONCATENATION OF FEATURES. WAV STANDS FOR SPEECH WAVEFORM. MEL IS SHORT FOR MEL-SPECTROGRAM. ENG, MAN, AND AVG INDICATE THE LANGUAGE OF THE CONVERTED SPEECH AND THEIR AVERAGE RESULTS

| | Experimental System | Configuration | | | | MCD | | | RMSE | | | WER/CER (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Input | Output | BNF Loss | ASR Loss | ENG | MAN | Avg | ENG | MAN | Avg | ENG | MAN | Avg |
| | Natural Source Speech | N.A. | | | | 8.71 | 8.94 | 8.83 | 18.08 | 19.93 | 19.01 | 8.21 | 3.75 | 5.98 |
| 1) | BNF2Wav | BNF | Wav | × | × | 8.77 | 9.01 | 8.89 | 13.24 | 13.41 | 13.33 | 21.66 | 17.83 | 19.75 |
| 2) | BNF2Wav($F0$) | BNF$\bigoplus F0$ | Wav | × | × | 8.69 | 8.81 | 8.75 | 13.19 | 13.17 | 13.18 | 21.68 | 17.77 | 19.73 |
| 3) | BNF2Mel-BNF | BNF | Mel | ✓ | × | 8.71 | 8.78 | 8.75 | 12.73 | 12.89 | 12.81 | 15.46 | 10.01 | 12.74 |
| 4) | BNF2Wav-BNF | BNF | Wav | ✓ | × | 7.85 | 7.96 | 7.91 | 12.52 | 12.40 | 12.46 | 12.10 | 9.98 | 11.04 |
| 5) | BNF2Wav-ASR | BNF | Wav | × | ✓ | 8.66 | 8.63 | 8.65 | 12.55 | 12.61 | 12.58 | 11.06 | 9.25 | 10.16 |
| 6) | BNF2Wav-BNF-ASR | BNF | Wav | ✓ | ✓ | 8.01 | 8.24 | 8.13 | 12.46 | 12.38 | 12.42 | 11.33 | 9.02 | 10.18 |
| 7) | BNF2Wav-BNF-ASR($F0$) | BNF$\bigoplus F0$ | Wav | ✓ | ✓ | 7.96 | 7.99 | 7.98 | 12.49 | 12.31 | 12.40 | 11.28 | 9.13 | 10.21 |

TABLE V
THE BNF2MEL CONVERSION MODEL CONFIGURATIONS WITH AN ENCODER AND A DECODER. FC IS THE FULLY CONNECTED LAYER, CONV-K-C-RELU DENOTES THE CONVOLUTION WITH WIDTH K, C OUTPUT CHANNELS, AND RELU ACTIVATION

| | |
|---|---|
| Encoder Pre-Net | FC-256-ReLU→Dropout(0.5) |
| Encoder CBHG | Conv1D<br>bank: K=8, conv-k-128-ReLU |
| | Max pooling<br>stride=1, width=2 |
| | Conv1D projections<br>conv-2-256-BactchNorm1D |
| | Highway Net<br>4 layers of FC-128-ReLU |
| | Bidirectional GRU<br>1 layer of 128 units |
| Decoder Pre-Net | FC-128-ReLU→Dropout(0.5)<br>→FC-128-ReLU→Dropout(0.5) |
| Decoder RNN | 1 layer of 256 GRU units |
| Decoder Post-Net | Conv1D projections<br>conv-4-256-BatchNorm1D-tanh |

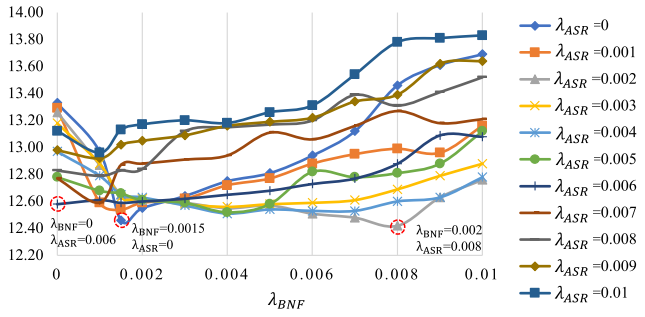


Fig. 5. The average RMSEs (dB) by varying the coefficient $\lambda_{BNF}$ and $\lambda_{ASR}$ for the BNF loss and ASR loss in (3). The selected combination of the $\lambda_{BNF}$ and $\lambda_{ASR}$ achieves the optimal RMSEs, which are shown in the red circle.

reconstruction loss becomes larger when one of the weights is larger than 0.01, resulting in worse performance. Therefore, we perform the experiments by varying the values from 0 to 0.01. The RMSE and WER/CER results are shown in Figs. 5 and 6. It can be found that both the WER/CER values consistently decrease with larger $\lambda_{BNF}$ and $\lambda_{ASR}$. However, Fig. 5 shows that the RMSE values first decrease with the increase of $\lambda_{BNF}$ or $\lambda_{ASR}$, but then increase. The final combination of the coefficients is determined by the optimal RMSEs.

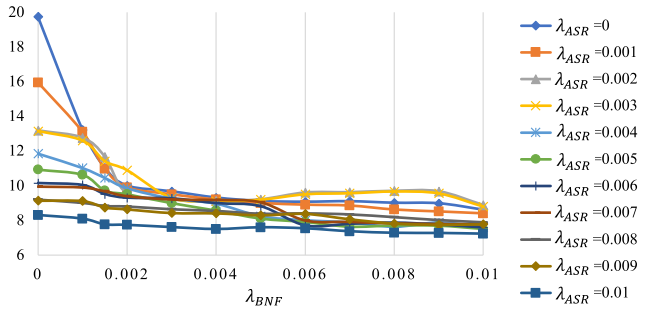


Fig. 6. The average WER/CER(%) by varying the coefficient $\lambda_{BNF}$ and $\lambda_{ASR}$ for the BNF loss and ASR loss in (3).

### C. Objective Evaluation

This study utilizes the Mel-Cepstral Distortion (MCD), the Root Mean Square Errors (RMSE), and the WER/CER as the objective metrics.

MCD is the spectral distance between the converted and target Mel-cepstrum features. It is calculated by the equation:

$$MCD[dB] = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{D}(\hat{M}_d - M_d)^2}, \qquad (4)$$

$\hat{M}_d$ and $M_d$ are the $d^{\text{th}}$ coefficients of the corresponding converted and target Mel-cepstrum.

RMSE represents the distortion of the converted speech waveform against the reference target speech. Particularly, the RMSE values between converted and reference utterances in frequency domain are calculated using the following equation:

$$RMSE[dB] = \sqrt{\frac{1}{K}\sum_{j=1}^{K}\left(20\log_{10}\frac{|\hat{F}_j|}{|F_j|}\right)^2}, \qquad (5)$$

where $K$ is the number of frequency bands, $|\hat{F}_j|$ and $|F_j|$ are the corresponding magnitude values of the short-time Fourier transform of the converted and reference speech waveform at the $j$-th frequency bin. $|\cdot|$ denotes the absolute operation.

The WER/CER with respect to the source speech, which is calculated following the same steps as in Section II-A, is the proxy of the speech intelligibility.

The average MCD, RMSE and WER/CER values of all experimental systems can be found in Table IV. For all metrics, the lower the values, the better the quality and intelligibility.

*1) Effect of End-to-End Optimization:* The only difference between BNF2Mel-BNF and BNF2Wav-BNF is that BNF2Wav-BNF optimizes the conversion model using the losses from the waveform directly in an end-to-end fashion, while BNF2Mel-BNF uses the Mel-spectrogram as an intermediate feature. With the end-to-end optimization, BNF2Wav-BNF achieves a reduction of MCD over BNF2Mel-BNF from 8.75 dB to 7.91 dB, a reduction of RMSE from 12.81 dB to 12.46 dB, and a reduction of WER/CER from 12.74% to 11.04%. The results suggest that both speech quality and intelligibility benefit from the end-to-end optimization. They also confirm that the BNF-to-Waveform framework, which performs end-to-end optimization, is an optimal choice for the conversion model. We will only use the BNF-to-Waveform framework in the remaining comparisons.

*2) Effect of Linguistic Losses:* With BNF2Wav as a baseline reference, it is observed that BNF2Wav-BNF, BNF2Wav-ASR and BNF2Wav-BNF-ASR reduce the MCD from 8.89 dB to 7.91 dB, 8.65 dB, and 8.13 dB, respectively; and the RMSE from 13.33 dB to 12.46 dB, 12.58 dB, and 12.42 dB, respectively. Based on the WER/CER metric, BNF2Wav-BNF, BNF2Wav-ASR and BNF2Wav-BNF-ASR reduce the WER/CER from 19.75% to 11.04%, 10.16%, and 10.18%, respectively. ASR loss also leads to lower WER/CER, as expected, because the ASR loss corresponds to the WER/CER objective function. The objective results suggest that both linguistic losses can help the conversion model to reduce distortion and pronunciation errors in the converted speech. This is a strong indicator that the linguistic losses can improve the subjective quality and intelligibility, which will be validated in the next section.

*3) Effect of Additional F0 Input:* The comparison of BNF2Wav and BNF2Wave($F0$) shows that the inclusion of $F0$ reduces the MCDs from 8.77 dB and 9.01 dB to 8.69 dB and 8.81 dB for English and Mandarin, respectively. Similarly, the RMSEs are also reduced from 13.24 dB and 13.41 dB to 13.19 dB and 13.17 dB for English and Mandarin, respectively. This indicates that the use of prosody information helps to produce a speech waveform with lower distortion. As expected, the RMSE reduction is larger in Mandarin than in English. This is because Mandarin is a tonal language and $F0$ is an important factor in intelligibility (i.e., the same pronunciation with a different tone has a different meaning). The results also show that $F0$ has almost no impact to the WERs/CERs. We argue that the influence of $F0$ on the WER/CER metric is minimal because the ASR model has already modeled the variation in accent and pronunciation.

To summarize, the BNF-to-Waveform framework performing end-to-end optimization results in lower distortion and WER/CER. Besides, linguistic losses considerably reduce MCD, RMSE and WER/CER. Both can be combined into one framework to further improve the performance. In addition, as expected, using $F0$ as part of the model input can slightly reduce RMSE, but not WER/CER.

### D. Subjective Evaluation

We further conducted subjective listening tests, including a MUSHRA test for speech quality and a speaker similarity test. 20 listeners who are proficient in both English and Mandarin Chinese participated in the tests.[5] Each listener evaluated 10 English and 10 Mandarin utterances randomly selected from 80 converted English and 80 converted Mandarin speech samples, respectively.

*MUSHRA test for speech quality:* In each query, listeners judged with converted samples from different XVC systems with the same speech content, and the reference natural speech from the target speaker was also included. In total, each query contained 8 speech samples arranged in random order. Listeners were instructed to rate the speech quality of each sample between 0 and 100. A higher score corresponds to better speech quality.

*Speaker similarity test:* To assess the speaker similarity of the converted samples with a reference audio, the same/different paradigm from VCC 2018 [63] was used. In each query, listeners were presented with the source speech, target speech, reference speech and 7 converted speech from different XVC systems. The source speech and the target speech are from the source and target speakers and are in different languages. The target speech and the reference speech are from the same target speaker but in different languages. The reference speech is used as a benchmark. During the experiment, each subject was asked to listen to the reference speech first and then to the other speech samples in random order. Then they had to decide whether this sample was from the same speaker as the reference speech by focusing only on the identity of the speaker. The decision was made on a scale of four, namely 'Different (sure), Different (not sure), Same (not sure), Same (sure)' [63].

Fig. 7 presents the MUSHRA score distributions of each XVC system. A higher median value of the distribution represents better speech quality. Fig. 8 presents the speaker similarity results. A higher percentage of 'Same (not sure)' and 'Same (sure)' together suggested a higher similarity to the desired target speaker. We will summarize the findings in the following sections.

*1) Evaluation of End-to-End Optimization:* Fig. 7 shows that the MUSHRA score of BNF2Wav-BNF is slightly higher than that of BNF2Mel-BNF. This indicates that converting BNFs directly to the waveform is better than using the Mel-spectrogram as an intermediate feature in terms of speech quality. The performance improvement is more obvious in Mandarin, which is in line with our expectations since the proposed linguistic losses can regulate the tonal information in Mandarin Chinese.

Fig. 8 shows that the total percentage of 'Same (not sure)' and 'Same (sure)' of the BNF2Mel-BNF system is much lower than that of the BNF2Wav-BNF. This means that using the BNF-to-Waveform XVC framework with end-to-end optimization is much better at capturing the identity of the target speaker than the BNF to Mel-spectrogram framework. The results confirm that using the BNF-to-Waveform framework is fairly effective for the XVC task. The subjective results are consistent with the objective results.

*2) Evaluation of Linguistic Losses:* From Fig. 7, we can observe that the systems integrated with linguistic losses, including

[5]Speech samples are available at https://vcsamples.github.io/taslp2022

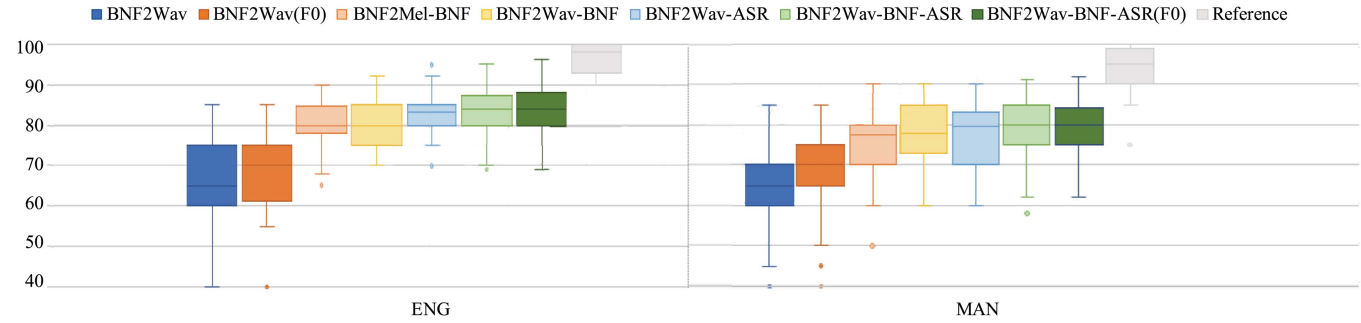


Fig. 7. Speech quality MUSHRA test scores. ENG and MAN indicate the languages of the converted speech. The target natural speech is used as the reference.

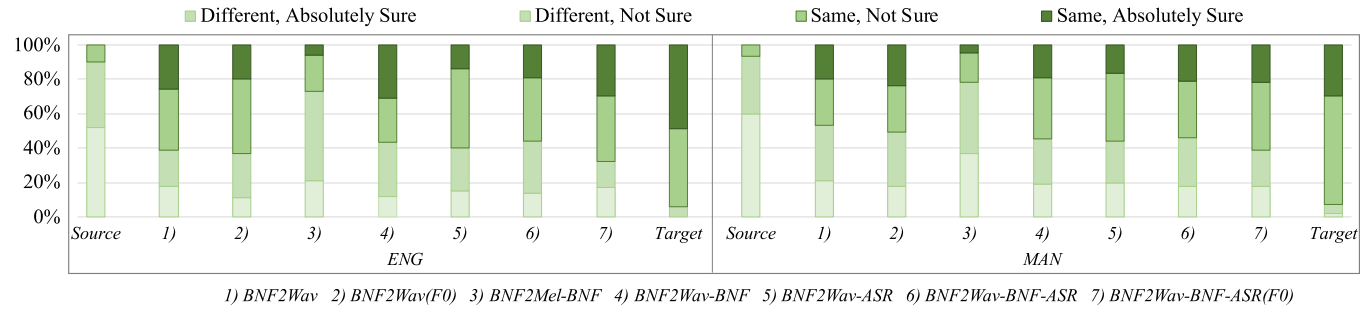


Fig. 8. Speaker similarity test results of all XVC systems. *Source* and *Target* denote the natural source speech and target speech. A higher percentage of 'Same (not sure)' and 'Same (sure)' together suggests a higher similarity to *Target*. *ENG* and *MAN* indicate the languages of the converted speech.

BNF2Mel-BNF, BNF2Wav-BNF, BNF2Wav-ASR, BNF2Wav-BNF-ASR, and BNF2Wav-BNF-ASR($F0$), significantly outperform BNF2Wav and BNF2Wav($F0$) by a large margin. This is strong evidence to demonstrate the effectiveness of the proposed linguistic losses in terms of speech quality and is consistent with the objective evaluations discussed in Section IV-C2.

By comparing BNF2Wav with BNF2Wav-BNF, BNF2Wav-ASR, and BNF2Wav-BNF-ASR in Fig. 8, we do not observe noticeable performance differences in English. However, we find that BNF2Wav-BNF, BNF2Wav-ASR, and BNF2Wav-BNF-ASR perform slightly better than BNF2Wav in the case of Mandarin. This means that the proposed linguistic losses can improve not only the quality but also the speaker similarity, especially for tonal languages.

*3) Evaluation on the Use of $F0$:* In Fig. 7, BNF2Wav($F0$) performs much better than BNF2Wav for both English and Mandarin. While BNF2Wav-BNF-ASR($F0$) and BNF2Wav-BNF-ASR receive a similar number of preferences among listeners. It suggests that using $F0$ may be beneficial to the XVC system. However, when system performance approaches realistic speech, whether using the additional $F0$ do not further impact the speech quality.

As far as speaker similarity is concerned, Fig. 8 presents that BNF2Wav($F0$) outperforms BNF2Wav and BNF2Wav-BNF-ASR($F0$) also outperforms BNF2Wav-BNF-ASR. This suggests that using $F0$ favors the XVC system in generating speech samples that are close to the target speaker's voice.

To summarize, we have the consistent observations in the subjective evaluations as that in the objective evaluations. The proposed linguistic losses not only improve speech quality but also speaker similarity. The overall results demonstrate that integrating the proposed linguistic losses into the BNF-to-Waveform framework successfully achieves a boosted cross-lingual voice conversion performance.

## V. Conclusion

This paper focuses on the speech intelligibility problem (i.e., foreign accent) in cross-lingual voice conversion. First, we investigate the speech intelligibility of the converted speech using VCC 2020 submitted systems and our in-house system. We confirm that cross-lingual voice conversion will produce converted speech with degraded speech intelligibility, and hypothesize that this is due to the different articulation of the languages. We then propose two linguistic losses to address the issue, one loss using the phonetic information at the frame level and the other loss using the character sequence, which is accent-independent. Cross-lingual voice conversion experiments between English and Mandarin have successfully demonstrated that both linguistic losses contribute to performance gains in terms of speech intelligibility. Therefore, the proposed framework facilities the cross-lingual voice conversion system to generate highly intelligible speech samples.

The focus of this work is cross-lingual voice conversion, where we assume high-quality recordings from the target speaker. Hence, text can be easily obtained and used to guide the voice conversion model training. If the ground-truth text is not provided, the ASR loss can be replaced with losses similar to the BNF loss in a self-supervised manner. In the future, we will investigate this direction with self-supervised techniques.

## REFERENCES


[1] M. Abe, K. Shikano, and H. Kuwabara, "Cross-language voice conversion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1990, pp. 345–348.

[2] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," *J. Inf. Process. J.*, vol. 43, no. 7, pp. 2177–2185, 2002.

[3] J. Dines et al., "Personalising speech-to-speech translation: Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis," *Comput. Speech Lang.*, vol. 27, no. 2, pp. 420–437, 2013.

[4] O. Turk, "Cross-lingual voice conversion," Bogaziçi University, Türkiye, vol. 3, 2007.

[5] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *Proc. INTERSPEECH*, 2016, pp. 2468–2472.

[6] M. J. Munro, "Foreign accent and speech intelligibility," *Phonology Second Lang. Acquisition*, vol. 5, pp. 193–218, 2008.

[7] X. Xie and C. A. Fowler, "Listening with a foreign-accent: The interlanguage speech intelligibility benefit in mandarin speakers of english," *J. Phonetics*, vol. 41, no. 5, pp. 369–378, 2013.

[8] T. M. Derwing and M. J. Munro, "Accent, intelligibility, and comprehensibility: Evidence from four l1s," *Studies Second Lang. Acquisition*, vol. 19, no. 1, pp. 1–16, 1997.

[9] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Lang. Learn.*, vol. 45, no. 1, pp. 73–97, 1995.

[10] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[11] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Proc. INTERSPEECH*, 2015, pp. 1468–1472.

[12] Y. Wang et al., "Transformer-based acoustic modeling for hybrid speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6874–6878.

[13] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 976–989, 2020.

[14] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1878–1889.

[15] X. Tian, J. Wang, H. Xu, E.-S. Chng, and H. Li, *Average Modeling Approach to Voice Conversion With Non-Parallel Data*. New York, NY, USA: Odyssey, 2018, pp. 227–232.

[16] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-VC2: Improved cyclegan-based non-parallel voice conversion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6820–6824.

[17] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5210–5219.

[18] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[19] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2009, pp. 3893–3896.

[20] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 922–931, Jul. 2010.

[21] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 4869–4873.

[22] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," in *Proc. INTERSPEECH*, 2016, pp. 287–291.

[23] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.

[24] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams," in *Proc. INTERSPEECH*, 2019, pp. 2843–2847.

[25] J.-X. Zhang et al., "Voice conversion by cascading automatic speech recognition and text-to-speech synthesis with prosody transfer," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 121–125.

[26] H. Kawanami, "GMM-based voice conversion applied to emotional speech synthesis," in *Proc. INTERSPEECH*, Sep. 2003, pp. 2401–2404.

[27] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, 2020, pp. 230–237.

[28] K. Yanagisawa and M. Huckvale, "A phonetic assessment of cross-language voice conversion," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 593–596.

[29] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 5120–5123.

[30] M. Davenport and S. J. Hannahs, *Introducing Phonetics and Phonology*. Evanston, IL, USA: Routledge, 2020.

[31] P. B. De Mareüil and B. Vieru-Dimulescu, "The contribution of prosody to the perception of foreign accent," *Phonetica*, vol. 63, no. 4, pp. 247–267, 2006.

[32] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2009, pp. 421–426.

[33] Z. Yi et al., "Voice conversion challenge 2020–intra-lingual semi-parallel and cross-lingual voice conversion," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 80–98.

[34] D. Erro and A. Moreno, "Frame alignment method for cross-lingual voice conversion," in *Proc. INTERSPEECH*, 2007, pp. 1969–1972.

[35] M. Zhang, J. Tao, J. Nurminen, J. Tian, and X. Wang, "Phoneme cluster based state mapping for text-independent voice conversion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2009, pp. 4281–4284.

[36] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 954–964, Jul. 2010.

[37] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6790–6794.

[38] C.-J. Chang, "Transfer learning from monolingual asr to transcription-free cross-lingual voice conversion," 2020, *arXiv:2009.14668*.

[39] L.-J. Liu et al., "Non-parallel voice conversion with autoregressive conversion model and duration adjustment," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 126–130, doi: 10.21437/VCC_BC.2020-17.

[40] Q. Ma, R. Liu, X. Wen, C. Lu, and X. Chen, "Submission from SRCB for voice conversion challenge 2020," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 131–135, doi: 10.21437/VCC_BC.2020-18.

[41] X. Tian et al., "The NUS & NWPU system for voice conversion challenge 2020," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 170–174.

[42] L. Zheng, J. Tao, Z. Wen, and R. Zhong, "CASIA voice conversion system for the voice conversion challenge 2020," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 136–139, doi: 10.21437/VCC_BC.2020-19.

[43] M. Wester and H. Liang, "The EMIME mandarin bilingual database," The University of Edinburgh, Edinburgh, Scotland, Tech. Rep. EDI-INF-RR-1396, 2011.

[44] N. Kalchbrenner et al., "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2415–2424.

[45] J. Johnson, A. Alahi, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[46] Y. Zhou, X. Tian, and H. Li, "Multi-task waveRNN with an integrated architecture for cross-lingual voice conversion," *IEEE Signal Process. Lett.*, vol. 27, pp. 1310–1314, 2020.

[47] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel waveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6199–6203.

[48] Y. Zhou, X. Tian, E. Yılmaz, R. K. Das, and H. Li, "A modularized neural network with language-specific output layers for cross-lingual voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 160–167.

[49] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," in *Proc. Interspeech*, 2016, pp. 322–326.
[50] Z. Yao et al., "WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*, 2021, pp. 4054–4058.
[51] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 5206–5210.
[52] "Aidatatang 200Zh, a free Chinese Mandarin speech corpus," Beijing DataTang Technology Co. Ltd.
[53] B. Hui, D. Jiayu, N. Xingyu, W. Bengu, and Z. Hao, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. 20th Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment*, 2017, pp. 1–5.
[54] "MAGICDATA Mandarin Chinese read speech corpus," Magic Data Technology Co., Ltd. 2019. [Online]. Available: http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101
[55] "PrimeWords chinese corpus set 1," Primewords Information Technology Co., Ltd. 2018. [Online]. Available: https://www.primewords.cn
[56] "ST-CMDS-20170001 1 free ST Chinese Mandarin corpus," Surfingtech.
[57] D. Wang and X. Zhang, "THCHS-30: A free chinese speech corpus," 2015, *arXiv:1512.01882*.
[58] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.
[59] Y. Zhou, X. Tian, and H. Li, "Language agnostic speaker embedding for personalized cross-lingual speech generation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3427–3439, 2021.
[60] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: Transforming mandarin ASR research into Industrial scale," 2018. *arXiv:1808.10583*.
[61] C. Veaux et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.
[62] Y. Zhou, X. Tian, Z. Wu, and H. Li, "Cross-lingual voice conversion with a cycle consistency loss on linguistic representation," in *Proc. Interspeech*2021, pp. 1374–1378.
[63] J. Lorenzo-Trueba et al., "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. INTERSPEECH*, 2018, pp. 195–202.